# Analysis of breast cancer related gene expression using natural splines and the Cox proportional hazard model to identify prognostic associations

Bas Kreike, Guus Hart, Harry Bartelink, Marc J. Vijver

Breast Cancer Res Treat (2010) 122:711–720
DOI 10.1007/s10549-009-0588-6

PRECLINICAL STUDY

# Analysis of breast cancer related gene expression using natural splines and the Cox proportional hazard model to identify prognostic associations

**Bas Kreike · Guus Hart · Harry Bartelink · Marc J. van de Vijver**

Received: 15 September 2009 / Accepted: 8 October 2009 / Published online: 27 October 2009


**Abstract** Many studies correlating gene expression data to clinical parameters assume a linear increase or decrease of the clinical parameter under investigation with the expression of a gene. We have studied genes encoding important breast cancer-related proteins using a model for survival-type data that is based on natural splines and the Cox proportional hazard model, thereby removing the linearity assumption. Expression data of 16 genes were studied in relation to metastasis-free probability in a cohort of 295 consecutive breast cancer patients treated at The Netherlands Cancer Institute. The independent predictive power for disease outcome of the 16 individual genes was tested in a multivariable model with known clinical and pathological risk factors. There is a linear relationship between increasing expression and a higher or lower hazard for distant metastasis for *ESR1*, *ERBB4*, *VEGF*, *CCNE2*, *EZH2*, and *UPA*; for *ERBB2*, *ERBB3*, *CCND1*, *CCNE1*, *EED*, *CXCR4*, *CCR7*, *SDF1*, and *PAI1* there is no clear increase or decrease; and for *EGFR* there seems to be a non-linear relation. Multivariable analysis showed that the 70-gene prognosis profile outperforms all the other variables in the model (hazard-rate 5.4, 95% CI 2.5–11.7; $P = 0.000018$). *EGFR*-expression seems to have a non-linear relation with disease outcome, indicating that lower but also higher expression of *EGFR* are associated with worse outcome compared to intermediate expression levels; the other genes show no or a linear relation.

**Keywords** Breast cancer · Microarray gene expression data · Natural spline analysis · Cox proportional hazard analysis · Distant metastasis-free probability

## Introduction

Gene expression profiling by microarray analysis in human cancer has proven to be a powerful tool to identify novel prognostic and predictive factors. Several studies have used microarray analysis to generate expression profiles predicting disease outcome, using unsupervised hierarchical clustering and supervised methods of analysis [1–10]. Until now, little attention is given to the exact nature of the relation between expression levels and outcome such as cancer recurrence. Often, implicitly or explicitly it is assumed that the outcome measure increases or decreases in a linear fashion with the gene expression level.

Prior to the emergence of microarray analysis, many studies of the prognostic significance of the expression level of single genes or proteins have been performed in breast cancer. A problem when analyzing gene expression levels in relation to breast cancer survival is the categorization of patients. The categorization is usually based on subjectively chosen cut-off points in the distribution of the expression levels. It is often not stated how these choices were made and this may lead to non-optimal categories resulting in loss of statistical power and precision.

**Electronic supplementary material** The online version of this article (doi:10.1007/s10549-009-0588-6) contains supplementary material, which is available to authorized users.

B. Kreike · G. Hart · H. Bartelink
Divisions of Radiation Oncology and Experimental Therapy, The Netherlands Cancer Institute, Plesmanlaan 121, Amsterdam 1066 CX, The Netherlands

M. J. van de Vijver (✉)
Department of Pathology, Academic Medical Center, Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands
e-mail: m.j.vandevijver@amc.uva.nl

In this study we use an alternative approach to the analysis of the relation between microarray results and breast cancer survival data. Our model is based on natural splines [11] and the Cox proportional hazard model. The basic model of Cox assumes a linear relation between the logarithm of the hazard rate and the expression level of a marker or gene. In our approach we use the natural spline to model this relation more flexibly. A spline has to pass through a certain number of points ("knots") not lying on a straight line. Guidelines have been formulated for the number and position of the knots in regression analyses such as the proportional hazard analysis [11]. Furthermore the natural spline approach avoids categorization of expression levels or patients. Therefore, the natural spline technique is a powerful tool in the statistical analysis of microarray data.

We used the natural spline technique to analyze the association of the expression of 16 (breast) cancer-relevant genes, measured using microarray analysis, with disease outcome. We studied estrogen receptor alpha (*ESR1*), genes encoding for proteins of the ERBB-family (*EGFR*, *ERBB2*, *ERBB3*, and *ERBB4*), cyclin-family members (*CCND1*, *CCNE1*, and *CCNE2*), Polycomb-group family members (*EZH2*, *EED*) and several genes that are thought to play a role in the development of distant metastases (*VEGF*, *CXCR4*, *CCR7*, *SDF1*, *PAI1*, and *UPA*). We also adjusted their individual prognostic power for confounding by other known clinical prognostic factors (e.g., clinical tumor stage, age, pathologic tumor characteristics, etc.) and the 70-genes prognostic profile previously identified by our group [1, 12].

## Materials and methods

Detailed patient information has been described previously [12]. Briefly, tumors from 295 women with breast cancer treated between 1984 and 1995 were selected from the fresh-frozen-tissue bank of The Netherlands Cancer Institute according to the following criteria: primary invasive breast carcinoma, <5 cm in diameter at pathological examination; age at diagnosis ≤52 years; and negative history of previous cancer except non-melanoma skin cancers. All patients received modified radical mastectomy or breast-conserving treatment, including dissection of the axillary lymph nodes and radiotherapy if indicated. Among the 295 patients, 151 had lymph-node-negative disease and 144 had lymph-node-positive disease. Ten of the 151 patients with lymph-node-negative disease and 120 of the 144 who had lymph-node-positive disease had received adjuvant chemotherapy ($n = 90$), hormonal therapy ($n = 20$), or both ($n = 20$). Patients were assessed at least annually after completion of therapy. Median duration of follow-up was 7.8 years (range, 0.05–18.3) for the 207 patients without metastasis as first event and 2.7 years (range, 0.3–14.0) for the 88 patients with metastasis as first event. The median follow-up among all 295 patients was 6.7 years (range, 0.05–18.3).

Details on RNA isolation, labeling of complementary RNA, competitive hybridization of each tumor cRNA with pooled reference cRNA from all samples to 24,500 element oligonucleotide microarrays, and measurement of expression ratios were previously described [1]; and wherever possible we adhered to reporting Recommendations for Tumor Marker Prognostic Studies (REMARK) [13]. Expression levels were quantified as log10($R$), where $R$ is the ratio of the intensities of test and reference sample and log10 denotes the logarithm with base 10.

### Statistical analysis

Natural splines and the Cox proportional hazard model were used to study the relation of gene expression levels and distant metastasis free probability (DMFP). Concisely, natural splines consist of connected piecewise cubic polynomials, each defined on a separate range of the expression level. The boundaries between the ranges are called knots. Guidelines have been given for the choice of the number and position of the knots. We used five knots at the positions advocated by Harrell [11]: 5, 25, 50, 75, and 95% quantiles of the expression level. More detailed information on these calculations are given in Supplementary material 1.

Table 1 displays the $P$ values of: the natural spline analysis; non-linearity test, and linear test. The "non-linearity" and "linear relation" $P$ values are independent from each other; however, the "natural spline" $P$ value depends on both. We used the "linear relation" $P$ value for the relation between the expression level and DMFP unless the "non-linearity" $P$ value is <0.01; for these cases the "natural spline" $P$ value was used. This summarizing $P$ value is given in Table 1 under the heading "overall".

By definition, a single hazard ratio can only be given in the linear model. Hazard ratios per unit expression level based on the linear model are given in Table 2 including the 95% confidence interval (CI).

The relation between the gene expression level and DMFP was visualized in graphs of the hazard rate as a function of the gene expression ratio; the units in which they are expressed, ln(hazard ratio) per unit expression level, may be difficult to interpret intuitively. Therefore, we also created survival-type curves, but these curves necessitate categorization of the expression level. We used a categorization scheme intended to reflect the shape of the hazard rate function as clearly as possible by creating a possibly large number of categories, where the number and widths of the categories depended on the shape of hazard rate function (see the thin vertical lines in Figs. 1a–4a). In

**Table 1** Rate of metastases and gene expression: *P* values using different models/hypotheses

| Gene | Natural spline | Non-linearity | Linear relation | Overall | Categorization |
|---|---|---|---|---|---|
| ESR1 | 0.012 | 0.55 | 0.0008 | 0.0008 | 0.0007 |
| *EGFR* | 0.0065 | 0.0039 | 0.69 | 0.0065 | 0.0013 |
| *ERBB2* | 0.02 | 0.05 | 0.045 | 0.045 | 0.0008 |
| *ERBB3* | 0.46 | 0.7 | 0.15 | 0.15 | 0.052 |
| *ERBB4* | 0.0063 | 0.28 | 0.0011 | 0.0011 | 0.0004 |
| *VEGF* | 0.0006 | 0.44 | <0.0001 | <0.0001 | <0.0001 |
| *CCND1* | 0.6 | 0.42 | 0.97 | 0.97 | 0.073 |
| *CCNE1* | 0.016 | 0.084 | 0.02 | 0.02 | 0.0046 |
| *CCNE2* | 0.0014 | 0.59 | <0.0001 | <0.0001 | 0.0001 |
| *EZH2* | 0.0074 | 0.59 | 0.0005 | 0.0005 | 0.0002 |
| *EED* | 0.15 | 0.083 | 0.22 | 0.22 | 0.0094 |
| *CXCR4* | 0.21 | 0.15 | 0.57 | 0.57 | 0.0035 |
| *CCR7* | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| *SDF1* | 0.018 | 0.03 | 0.079 | 0.079 | 0.0016 |
| *PAI1* | 0.029 | 0.11 | 0.022 | 0.022 | 0.0012 |
| *UPA* | 0.0007 | 0.015 | 0.0053 | 0.0053 | <0.0001 |

the next step, adjacent categories were concatenated in a stepwise manner until further concatenation would result in a loss of information as measured by the Akaike Information Criterion (AIC) [14]. The complete process was performed in an automated way.

The remaining categories are shown by the thick vertical lines in the ln(hazard ratio) plots and the resulting survival curves are shown in Figs. 1b–4b. Nominal *P* values for the differences between the curves are given under the heading "categorization" in Table 1. These *P* values are calculated from the log-rank test, without taking into account the ordering of the categories ("nominal scale"). A large difference between the log-rank and overall *P* value indicates that differences in prognosis are overestimated by the Kaplan–Meier curves. Five- and 10-year metastasis-free percentages are given in Table 2.

In order to get an impression of the variability of the categorization process described above, 100 bootstrap samples of size 295 were created from the original data and the full process of categorization was repeated 100 times.

Normal probability-plots were used to identify outliers. The appearance of humps in the normal plot may indicate the existence of subpopulations.

Conclusions were generally formulated in terms of the strength of evidence. These are mainly based on *P* values from the linear or spline models using the following categorization: $P > 0.05$ = "no evidence"; $0.05 < P < 0.01$ = "not much evidence"; $0.01 < P < 0.001$ = "some evidence"; $0.001 < P < 0.0001$ = "evidence"; $P < 0.0001$ = "proof".

We tested the individual prognostic power of the selected genes in relation to known clinical and pathological risk factors for DMFP in a multivariable analysis. The 16 selected genes were tested in one model together with use of chemotherapy, hormonal therapy, type of surgery (mastectomy versus breast conserving surgery), tumor diameter; number of tumor involved lymphnodes; tumor grade, presence of vascular invasion, age, estrogen receptor status, and prognostic class assignment based on the 70-gene prognosis signature [1]. The use of systemic

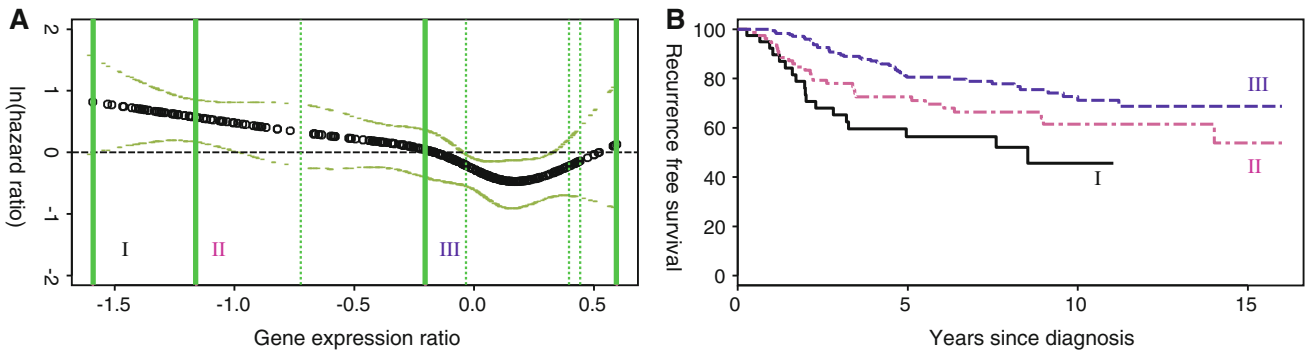


**Fig. 1** *ESR1* output. **a** Spline plot, ln(hazard ratio) and categorization for *ESR1*. The color-coded roman numbers indicate the groups that are compared in (**b**) and correspond to the curves with the same color. **b** Recurrence free interval based on the categorization of the spline plot

therapy, type of surgery, histological grade, presence of vascular invasion, and estrogen receptor status were used as nominal variables; all other variables were used in a linear fashion.

Patients with missing values for the relevant variables were omitted from the analysis. Also patients with extreme expression levels as identified in the Normal probability plots were excluded from analyses, which included the relevant gene. In the case of the multivariable analysis, this resulted in the use of 282 of the original 295 patients.

The statistical package S+© (Insightful Corp, Seattle) version 6.1–6.2 was used for the calculations and the graphs. This package implements the natural spline through a simple function. A user-written function for categorization can be obtained from one of the coauthors (G. Hart).

## Results

From a total of 24,500 genes we selected 16 genes for analysis. Table 1 shows the 16 genes of interest and their level of significance. Seven of the 16 genes show at least some level of evidence for an association with DMFP based on the spline model (overall $P$ value <0.01: *ESR1*, *EGFR*, *ERBB4*, *VEGF*, *CCNE2*, *EZH2*, and *UPA*). Of these 7 genes 6 show a significant linear relation between gene-expression level and hazard ratio (*ESR1*, *ERBB4*, *VEGF*, *CCNE2*, *EZH2*, and *UPA*). Only *EGFR* shows some evidence of a non-linear relation (non-linearity $P$ value = 0.0039). For the other nine genes there is no or not much evidence to have significant prognostic properties (overall $P$ value >0.01: *ERBB2*, *ERBB3*, *CCND1*, *CCNE1*, *EED*, *CXCR4*, *CCR7*, *SDF1*, and *PAI1*). Table 2 shows the hazard ratio (calculated for the linear model) for the individual genes and the categorization into groups defined by the gene-expression level cut-offs with 5 and 10 years metastases free probability.

The normal plot for *ESR1* suggests that the total population consists of two subpopulations of sizes of about 54 ($\log(R)$ less than $-1$) and 215 patients ($\log(R)$ greater than $-0.5$), respectively, with 26 patients in between (data not shown). These groups reflect the estrogen receptor (ER) status, the first group being the ER-negative patients and the second group the ER-positive, as has been shown previously [1, 12]. The natural spline analysis (Fig. 1a) provides evidence that higher expression levels of this marker are associated with lower rate of distant metastasis (5 year DMFP: low expression: 56%, intermediate expression: 72%, high expression: 81%, $P = 0.0008$). Figure 1a also shows a possible worse prognosis for the very high expression levels, this is, however, not captured by the categorization. The resulting categories roughly coincide with the subpopulations indicated by the normal plot. However, the bootstrap results indicate still considerable uncertainty about the number of categories and the position of the boundaries (data not shown).

The natural spline analysis for *EGFR* (Fig. 2a) shows a non-linear relation between gene-expression level and DMFP. A group of 163 patients have gene-expression ratios around zero and excellent survival rates of 84 and 75%, respectively, for 5 and 10 year DMFP. Patients with high expression ($n = 46$) or low expression ($n = 69$) of this gene have a poor prognosis with 5 and 10 year DMPF of, respectively, 67 and 50% for the patients with high expression of *EGFR* and 61 and 53% for the patients with low expression of *EGFR*.

According to our model there is proof that the rate of metastases increases with increasing levels of *VEGF* or *CCNE2*, $P < 0.0001$, with hazard rates of 4.2 and 4.5, respectively (Table 2; Fig. 3; Supplementary material 2-S1). There is also strong evidence suggesting that higher *EZH2* gene expression is associated with higher rates of distant metastasis, $P = 0.0005$ (Supplementary material 2-S2). The hazard rate plot suggests a rather sharp transition from low risk at expression levels below $-0.2$ to high risk at levels above 0.15. However, linearity is not ruled out, non-linear $P$ value = 0.59.

With a hazard ratio of 0.29 *ERBB4* shows evidence ($P = 0.0011$) that higher rates of distant metastases occur when this gene is expressed at low levels (Supplementary material 2-S3). Five year DMFP of 58 vs. 67–86% for low levels of ERBB4 expression versus the higher levels and 10 year DMFP of, respectively, 51 vs. 60–80% (Table 2).

*UPA* also seems to be associated with DMFP ($P = 0.0053$). Especially the patients with very high expression levels of *UPA* ($n = 11$) showed poor outcome as also reflected by the Kaplan–Meier curves (Supplementary material 2-S4), with 5 year DMFP of 27 vs. 73–78% for the patients with lower expression levels of *UPA* and 10 year DMFP of, respectively, 18 vs. 59–70% (Table 2).

The other nine genes (*ERBB2-3*, *CCND1-E1*, *EED*, *CXCR4*, *CCR7*, *SDF1*, and *PAI1*) were not shown to be associated with outcome based on our predefined $P$ value cut-off levels (Fig. 4; Supplementary material 2-S5–12).

The multivariable analysis showed that the 70-gene prognosis profile [1] outperformed the clinical risk factors and the individual gene expression ratios in predicting DMFP with a HR of 5.4 (95% CI 2.5–11.7; $P = 0.000018$); Table 3.

## Discussion

We have used natural splines and the Cox proportional hazard model to study the relation of gene expression levels and disease outcome. We have developed this model of

**Table 2** Hazard ratio (95% CI) per measurement unit for linear model and metastases-free percentages per category

| Gene | HR (95% CI) or Categories defined by gene expression levels | *N* | % free from distant metastases (SE) | |
|---|---|---|---|---|
| | | | 5-year | 10-year |
| ESR1 | HR: 0.54 (0.38–0.77) | | | |
| | *−1.591/−1.159* | 39 | 56% (8%) | 46% (10%) |
| | *−1.159/−0.203* | 79 | 72% (5%) | 61% (6%) |
| | *−0.203/0.596* | 177 | 81% (3%) | 71% (4%) |
| *EGFR* | HR: 0.63 (0.05–8.17) | | | |
| | *−0.388/−0.197* | 10 | 80% (13%) | 80% (13%) |
| | *−0.197/−0.076* | 69 | 61% (6%) | 53% (7%) |
| | *−0.076/0.046* | 163 | 84% (3%) | 74% (4%) |
| | *0.046/0.326* | 49 | 67% (7%) | 50% (10%) |
| *ERBB2* | HR: 1.57 (0.97–2.54) | | | |
| | *−1.188/−0.7* | 60 | 65% (6%) | 61% (7%) |
| | *−0.7/0.361* | 195 | 83% (3%) | 70% (4%) |
| | *0.361/0.928* | 40 | 54% (8%) | 51% (8%) |
| *ERBB3* | HR: 0.45 (0.15–1.37) | | | |
| | *−0.574/0.104* | 207 | 72% (3%) | 63% (4%) |
| | *0.104/0.444* | 83 | 82% (4%) | 72% (6%) |
| *ERBB4* | HR: 0.29 (0.14–0.63) | | | |
| | *−0.767/−0.228* | 78 | 58% (6%) | 51% (7%) |
| | *−0.228/0.009* | 88 | 80% (4%) | 60% (7%) |
| | *0.009/0.338* | 101 | 86% (3%) | 80% (5%) |
| | *0.338/0.525* | 26 | 67% (9%) | 67% (9%) |
| *VEGF* | HR: 4.2 (2.26–7.78) | | | |
| | *−0.65/−0.012* | 178 | 83% (3%) | 75% (4%) |
| | *−0.012/0.291* | 75 | 67% (6%) | 56% (7%) |
| | *0.291/0.89* | 42 | 59% (8%) | 44% (9%) |
| *CCND1* | HR: 1.01 (0.51–2) | | | |
| | *−0.982/−0.598* | 25 | 63% (10%) | 45% (13%) |
| | *−0.598/1.089* | 270 | 76% (3%) | 67% (3%) |
| *CCNE1* | HR: 7.22 (1.09–47.84) | | | |
| | *−0.572/−0.292* | 10 | 90% (9%) | 90% (9%) |
| | *−0.292/0.112* | 244 | 77% (3%) | 66% (4%) |
| | *0.112/0.467* | 37 | 55% (9%) | 49% (9%) |
| *CCNE2* | HR: 4.49 (2.11–9.55) | | | |
| | *−0.835/−0.179* | 122 | 87% (3%) | 76% (4%) |
| | *−0.179/0.096* | 99 | 73% (5%) | 66% (6%) |
| | *0.096/0.726* | 74 | 59% (6%) | 44% (7%) |
| *EZH2* | HR: 4.68 (2.06–10.63) | | | |
| | *−0.682/−0.124* | 100 | 89% (3%) | 76% (5%) |
| | *−0.124/0.132* | 118 | 71% (4%) | 69% (5%) |
| | *0.132/0.664* | 76 | 62% (6%) | 46% (7%) |
| *EED* | HR: 2.33 (0.52–10.4) | | | |
| | *−0.567/0.215* | 276 | 76% (3%) | 67% (3%) |
| | *0.215/0.578* | 17 | 53% (12%) | 40% (15%) |
| *CXCR4* | HR: 1.31 (0.54–3.18) | | | |
| | *−0.753/−0.18* | 92 | 82% (4%) | 75% (5%) |
| | *−0.18/0.145* | 144 | 68% (4%) | 55% (5%) |
| | *0.145/0.677* | 59 | 82% (5%) | 75% (7%) |

**Table 2** continued

| Gene | HR (95% CI) or Categories defined by gene expression levels | *N* | % free from distant metastases (SE) | |
|---|---|---|---|---|
| | | | 5-year | 10-year |
| *CCR7* | NA | | | |
| *SDF1* | HR: 0.29 (0.08–1.14) | | | |
| | *−0.513/−0.024* | 117 | 64% (5%) | 54% (6%) |
| | *−0.024/0.601* | 178 | 82% (3%) | 72% (4%) |
| *PAI1* | HR: 3.05 (1.23–7.58) | | | |
| | *−0.648/0* | 180 | 81% (3%) | 72% (4%) |
| | *0/0.775* | 115 | 67% (4%) | 56% (5%) |
| *UPA* | HR: 3.78 (1.36–10.46) | | | |
| | *−0.685/0.087* | 218 | 78% (3%) | 70% (4%) |
| | *0.087/0.434* | 66 | 73% (6%) | 59% (7%) |
| | *0.434/0.602* | 11 | 27% (13%) | 18% (12%) |

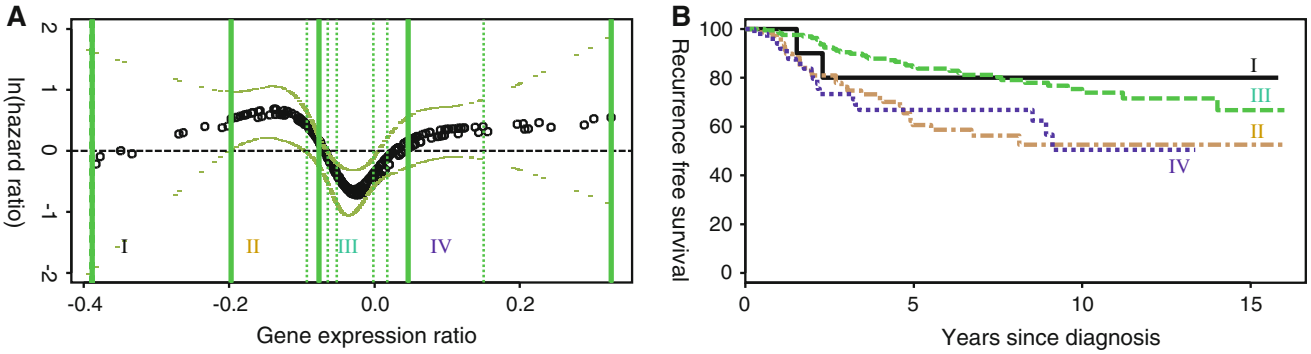


**Fig. 2** *EGFR* output. **a** Spline plot, ln(hazard ratio) and categorization for *EGFR*. The color-coded roman numbers indicate the groups that are compared in (**b**) and correspond to the curves with the same color. **b** Recurrence free interval based on the categorization of the spline plot

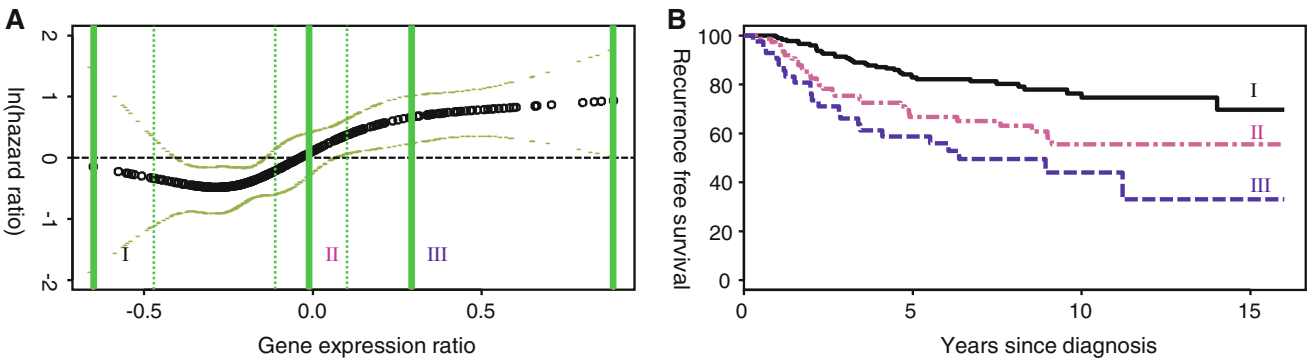


**Fig. 3** *VEGF* output. **a** Spline plot, ln(hazard ratio) and categorization for *VEGF*. The color-coded roman numbers indicate the groups that are compared in (**b**) and correspond to the curves with the same color. **b** Recurrence free interval based on the categorization of the spline plot

analysis to overcome the assumption that gene expression levels are linearly related to disease outcome. We analyzed the expression ratios of 16 genes (i.e., *ESR1*, *ERBB1*, *ERBB2*, *ERBB3*, *ERBB4*, *VEGF*, *CCND1*, *CCNE1*, *CCNE2*, *EZH2*, *EED*, *CXCR4*, *CCR7*, *SDF1*, *PAI1*, and *UPA*) in a cohort of 295 patients with long-term follow-up data. These genes were analyzed because of their known relation to breast cancer survival [15–26].

As expected from other studies *ESR1*, *EGFR*, *ERBB4*, *VEGF*, *CCNE2*, *EZH2*, and *UPA* are associated with distant metastasis free probability [15, 20, 21, 24–26]. We found that *EGFR* seems to be related to DMFP in a

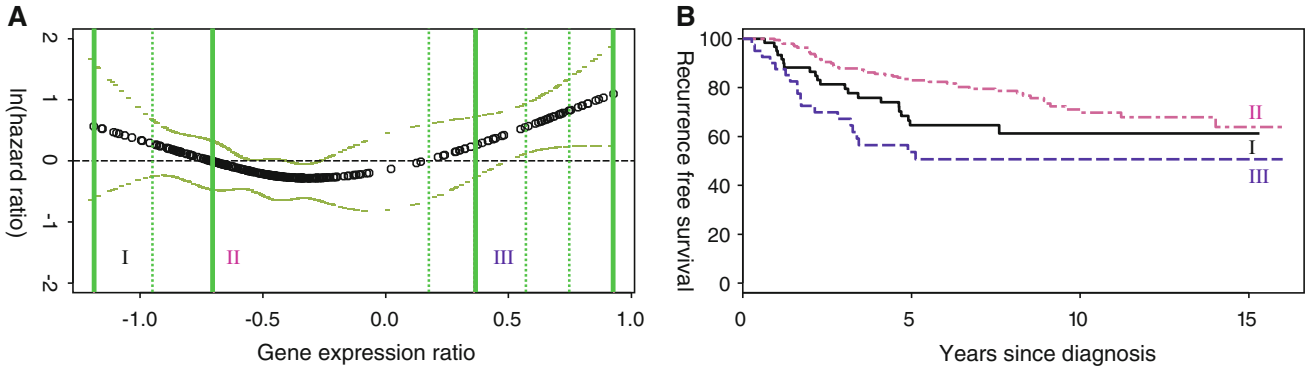


**Fig. 4** *ERBB2* output. **a** Spline plot, ln(hazard ratio) and categorization for *ERBB2*. The color-coded roman numbers indicate the groups that are compared in (**b**) and correspond to the curves with the same color. **b** Recurrence free interval based on the categorization of the spline plot

**Table 3** Multivariable analysis testing the prognostic power of the individual genes in relation to know clinical and pathological risk factors

| Variable | Hazard ratio | 95% confidence interval | *P* value |
|---|---|---|---|
| *ESR1* | 0.39 | 0.12–1.34 | 0.14 |
| *EGFR* | 0.24 | 0.01–4.35 | 0.34 |
| *ERBB2* | 1.12 | 0.68–1.85 | 0.65 |
| *ERBB3* | 2.14 | 0.30–15.3 | 0.45 |
| *ERBB4* | 0.72 | 0.22–2.33 | 0.59 |
| *VEGF* | 2.55 | 0.91–7.14 | 0.074 |
| *CCND1* | 1.14 | 0.36–3.57 | 0.82 |
| *CCNE1* | 1.03 | 0.07–16.3 | 0.98 |
| *CCNE2* | 0.68 | 0.18–2.53 | 0.56 |
| *EZH2* | 1.40 | 0.31–6.26 | 0.66 |
| *EED* | 4.48 | 0.72–27.9 | 0.11 |
| *CXCR4* | 0.17 | 0.04–0.72 | 0.016 |
| *CCR7* | 0.61 | 0.22–1.68 | 0.34 |
| *SDF1* | 1.51 | 0.20–11.6 | 0.69 |
| *PAI1* | 1.40 | 0.31–6.33 | 0.66 |
| *UPA* | 1.68 | 0.35–8.16 | 0.52 |
| 70-gene prognosis profile [1] | 5.40 | 2.50–11.7 | 0.000018 |
| Chemotherapy | 0.33 | 0.18–0.60 | 0.00025 |
| Hormonal therapy | 0.49 | 0.19–1.30 | 0.15 |
| Type of surgery | 1.51 | 0.88–2.59 | 0.13 |
| Tumor diameter | 1.04 | 1.01–1.07 | 0.0053 |
| Number of tumor involved lymphnodes | 1.15 | 1.05–1.25 | 0.0017 |
| Histological grade | 0.89 | 0.54–1.46 | 0.63 |
| Presence of vascular invasion | 1.32 | 0.97–1.79 | 0.075 |
| Age | 0.97 | 0.92–1.02 | 0.20 |
| Estrogen receptor status | 1.94 | 0.61–6.16 | 0.26 |

non-linear manner, meaning that tumors with high or low expression levels of *EGFR* have a poor prognosis compared to the tumors with an average expression. This has not been reported before and might be due to the fact that most researchers assume a linear relation between gene expression and disease outcome. In many studies high expression levels of *EGFR* is associated with poor survival [25, 27, 28]. When we would ignore the non-linear relation of *EGFR* and DMFP and analyze the present *EGFR*-data in a linear manner, we would not have identified *EGFR* as a marker for DMFP, since tumors with low expression levels and tumors with high expression levels of *EGFR* are not significant differently associated with DMFP.

We found a linear relation of *ESR1* and DMFP; this refutes previous reports that very high levels of ER are associated with poor outcome [10, 29], e.g., luminal B-like versus luminal A-like breast carcinomas. However, the shape of the spline plot (Fig. 1a) does indicate a slight increased risk at the high end of the expression level of *ESR1*, but this increase is too small to become significant.

Kleer et al. have reported on the association of *EZH2* expression with outcome in breast cancer, as an extension of similar findings in prostate cancer [5, 30]. They showed that elevated levels of *EZH2* expression were associated with poor survival in prostate cancer patients [30] and in breast cancer patients [15]. This is in agreement with our results that show that the chance of developing breast cancer metastases increases with increasing expression levels of the *EZH2* gene, following a sigmoid-like curve relation that can be regarded as essentially still linear. They compared tumors with high levels of EZH2-staining to tumors with low levels of EZH2-staining, using a subjective cutoff level chosen in advance of the actual analysis. Kleer et al. showed that EED, a protein known to form complexes with EZH2 to silence genes, was not expressed at high levels in invasive breast cancer tissue compared to normal breast tissue [15]. We did not observe a difference in expression level of *EED* between tumors with good or bad prognosis. We did identify a small subgroup ($n = 17$) of patients with high levels of *EED* (log10 expression ratio >0.22) that had worse DMFP than patients with lower expression levels (log-rank $P$ value = 0.0094), 5 and 10 year DMFP of 53 and 40% versus 76 and 67%, respectively (Supplementary material 2-S8). However, since the difference between the natural spline analysis and the forced categorization for the Kaplan–Meier analysis are rather large, we think that the log-rank $P$ value is an over-estimation of the reality.

For *ERBB2* (also known as *HER2/neu*) it is generally accepted that there is a non-linear relation between gene expression and disease outcome. It has been found that approximately 20% of all breast carcinomas are ERBB2-positive and this is associated with a worse breast cancer survival as compared to the ERBB2-negative patients. ERBB2-status is usually assessed using immunohistochemical techniques (combined with in situ hybridization). Surprisingly, the natural spline analysis of *ERBB2* (Fig. 4a) suggests a worse prognosis at both ends of the expression level distribution, but the $P$ values are still too large to allow a definite conclusion ($P = 0.045$).

For *VEGF* and *CCNE2* it is known that tumors progress more rapid and have a poor prognosis when these genes are over expressed [19, 24, 26, 31, 32]. This is in agreement with our results showing that higher expression levels of these genes are associated with poorer disease outcome in a linear fashion. The results of our analyses concerning these two genes had $P$ values <0.0001, representing a very significant association between gene expression and disease outcome.

UPA and PAI1 protein expression as assessed with ELISA assays have been identified as prognostic factors, also when tested in a randomized control trial [33]. It was found that low levels of UPA and PAI1 were associated with good breast cancer survival as compared to high levels of these proteins. We did not find strong evidence that *PAI1* is powerful enough to predict prognosis. *UPA* seems to be able to predict DMFP, however, the natural spline analysis clearly shows that this is particularly true for the tumors with very high expression levels of the gene. *UPA* did not seem to be able to differentiate in the large proportion of tumors with lower levels of UPA-expression (log10 expression ration <0.434). Five year DMFP was 73% for the group of tumors with an intermediate expression level versus 78% for the group of tumors with the lowest expression levels.

Several studies have suggested [23] that chemokines and their receptors (*CXCR4*, *CCR7*, and *SDF1*) are involved in the process of breast cancer metastasizing to the bone. Our analysis did not show a relationship of *CCR7* with the development of distant metastasis anywhere in the body (overall $P$ value = 1.00). Expression levels of *CXCR4* and *SDF1* were not found to have a relation with rate of distant metastasis (overall $P$ values, respectively, 0.57 and 0.079). This is in agreement with a study performed by Sun et al. in prostate cancer [34].

Although the expression of several individual genes is associated with DMFP in the univariable analysis, they loose their significance if tested in combination with the previously identified 70-gene prognosis profile in multivariable analysis.

We have found that the natural spline model can be easily applied to the standard analysis of the association between individual gene expression data and outcome in breast cancer. Other methods, notably other types of splines [11] or fractional polynomials [35] have comparable properties and the choice between these methods is somewhat subjective.

Natural spline and comparable approaches avoid categorization of expression levels. This is important as categorization generally results in loss of statistical power and leads to a categorization that may not be optimal. Our way of deriving the categories is data-driven and therefore may lead to more optimal cutoff points. The $P$ value based on this categorization may be quite biased as shown for instance by comparing it to the overall $P$ value for *EED* and *ERBB2* in Table 1. The natural spline-based $P$ value (Table 1) on the other hand is unbiased and can be used instead. Therefore, we would argue to base the statistical analysis itself on tools like natural splines, but to base presentation of results on categorization if necessary.

There are several potential confounding factors that may have influenced the results of our study. The sample size of our study is relatively small, which makes it unfeasible to analyze important subgroups, such as the group of patients with very high expression values of *ESR1* and the group with very low expression values of *ERBB2*. Furthermore, in the multivariable analysis we used all variables in a linear fashion, in order to quote hazard ratios for Table 3, consequently *EGFR* was also used linearly. Nevertheless, we think that the associations we have shown are solid despite these possible confounder.

In conclusion, the natural spline technique is a powerful tool to analyze the exact nature of the relation between gene expression levels and outcome of disease. It may also help in defining categories of patients with relatively homogeneous prognosis. In this way we found that *EGFR* may have a non-linear relation with disease outcome, a finding that cannot be revealed by most statistical analyses used to date. We found that at multivariable analysis the 70-gene prognosis profile outperformed the clinicopathological risk factors and the individual gene expression ratios in predicting distant metastasis free probability.

**Acknowledgment** This study was supported with a grant from the Dutch Cancer Society (NKB2002-2575).

## References

1. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–536
2. Ahr A, Holtrich U, Solbach C, Scharl A, Strebhardt K, Karn T et al (2001) Molecular classification of breast cancer patients by gene expression profiling. J Pathol 195:312–320
3. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE et al (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 8:816–824
4. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M et al (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406:536–540
5. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K et al (2001) Delineation of prognostic biomarkers in prostate cancer. Nature 412:822–826
6. Kihara C, Tsunoda T, Tanaka T, Yamana H, Furukawa Y, Ono K et al (2001) Prediction of sensitivity of esophageal tumors to adjuvant chemotherapy by cDNA microarray analysis of gene-expression profiles. Cancer Res 61:6474–6479
7. Perou CM, Sorlie T, Eisen MB, van de Rijin M, Jeffrey SS, Rees CA et al (2000) Molecular portraits of human breast tumours. Nature 406:747–752
8. Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME et al (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415: 436–442
9. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI et al (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 346:1937–1947
10. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H et al (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 98:10869–10874
11. Harrell FE (ed) (2001) General aspects of fitting regression models. In: Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer, New York, pp 23–24
12. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW et al (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347:1999–2009
13. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM (2005) REporting recommendations for tumour MARKer prognostic studies (REMARK). Eur J Cancer 41:1690–1696
14. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F (eds) Second international symposium on information theory. Akademiai Kiado, Budapest, pp 267–281
15. Kleer CG, Cao Q, Varambally S, Shen R, Ota I, Tomlins SA et al (2003) EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. Proc Natl Acad Sci USA 100:11606–11611
16. Rudolph P, Kuhling H, Alm P, Ferno M, Baldetorp B, Olsson H et al (2003) Differential prognostic impact of the cyclins E and B in premenopausal and postmenopausal women with lymph node-negative breast cancer. Int J Cancer 105:674–680
17. Coradini D, Daidone MG (2004) Biomolecular prognostic factors in breast cancer. Curr Opin Obstet Gynecol 16:49–55
18. Toi M, Bando H, Kuroi K (2000) The predictive value of angiogenesis for adjuvant therapy in breast cancer. Breast Cancer 7: 311–314
19. Kushlinskii NE, Gershtein ES (2002) Role of vascular endothelial growth factor during breast cancer. Bull Exp Biol Med 133:521–528
20. Ali S, Coombes RC (2000) Estrogen receptor alpha in human breast cancer: occurrence and significance. J Mammary Gland Biol Neoplasia 5:271–281
21. Harbeck N, Kates RE, Gauger K, Willems A, Kiechle M, Magdolen V et al (2004) Urokinase-type plasminogen activator (uPA) and its inhibitor PAI-I: novel tumor-derived factors with a high prognostic and predictive impact in breast cancer. Thromb Haemost 91:450–456
22. Look M, van Putten W, Duffy M, Harbeck N, Christensen IJ, Thomssen C et al (2003) Pooled analysis of prognostic impact of uPA and PAI-1 in breast cancer patients. Thromb Haemost 90: 538–548
23. Muller A, Homey B, Soto H, Ge N, Catron D, Buchanan ME et al (2001) Involvement of chemokine receptors in breast cancer metastasis. Nature 410:50–56
24. Keyomarsi K, Tucker SL, Buchholz TA, Callister M, Ding Y, Hortobagyi GN et al (2002) Cyclin E and survival in patients with breast cancer. N Engl J Med 347:1566–1575
25. Koutras AK, Kalogeras KT, Dimopoulos MA, Wirtz RM, Dafni U, Briasoulis E et al (2008) Evaluation of the prognostic and predictive value of HER family mRNA expression in high-risk early breast cancer: a Hellenic Cooperative Oncology Group (HeCOG) study. Br J Cancer 99:1775–1785
26. Gomez-Esquer F, Agudo D, Martinez-Arribas F, Nunez-Villar MJ, Schneider J (2004) mRNA expression of the angiogenesis markers VEGF and CD105 (endoglin) in human breast cancer. Anticancer Res 24:1581–1585
27. Navolanic PM, Steelman LS, McCubrey JA (2003) EGFR family signaling and its association with breast cancer development and resistance to chemotherapy (Review). Int J Oncol 22:237–252
28. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z et al (2004) Immunohistochemical and clinical characterization of the

basal-like subtype of invasive breast carcinoma. Clin Cancer Res 10:5367–5374
29. Dai H, van't Veer L, Lamb J, He YD, Mao M, Fine BM et al (2005) A cell proliferation signature is a marker of extremely poor outcome in a subpopulation of breast cancer patients. Cancer Res 65:4059–4066
30. Varambally S, Dhanasekaran SM, Zhou M, Barrette TR, Kumar-Sinha C, Sanda MG et al (2002) The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature 419:624–629
31. Sieuwerts AM, Look MP, Meijer-van Gelder ME, Timmermans M, Trapman AM, Garcia RR et al (2006) Which cyclin E prevails as prognostic marker for breast cancer? Results from a retrospective study involving 635 lymph node-negative breast cancer patients. Clin Cancer Res 12:3319–3328
32. Desmedt C, Ouriaghli FE, Durbecq V, Soree A, Colozza MA, Azambuja E et al (2006) Impact of cyclins E, neutrophil elastase and proteinase 3 expression levels on clinical outcome in primary breast cancer patients. Int J Cancer 119:2539–2545
33. Janicke F, Prechtl A, Thomssen C, Harbeck N, Meisner C, Untch M et al (2001) Randomized adjuvant chemotherapy trial in high-risk, lymph node-negative breast cancer patients identified by urokinase-type plasminogen activator and plasminogen activator inhibitor type 1. J Natl Cancer Inst 93:913–920
34. Sun YX, Wang J, Shelburne CE, Lopatin DE, Chinnaiyan AM, Rubin MA et al (2003) Expression of CXCR4 and CXCL12 (SDF-1) in human prostate cancers (PCa) in vivo. J Cell Biochem 89:462–473
35. Royston P (2000) A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. Stat Med 19: 1831–1847